科技大数据增值丰富化方法研究与工具研发*

孔贝贝[1]　谢　靖[1,2]　钱　力[1,2]　常志军[1,2]　吴振新[1,2]
[1](中国科学院文献情报中心　北京 100190)
[2](中国科学院大学经济与管理学院图书情报与档案管理系　北京 100190)

摘要: 【**目的**】解决科技大数据数据源分散、质量不高、内容单薄等问题。【**方法**】采用数据清洗、实体对齐、实体字段融合、冲突检测等增值计算方法，设计开发一套科技大数据增值丰富化的工具。【**结果**】通过本文研发的丰富化工具，在人员、机构、会议、期刊实体及实体关系层面实现实体数据对齐，实体字段内容增加 5-10 倍，实体分析维度提升 2-3 倍。【**局限**】增值数据的及时性、规范性需要结合服务需求在实际应用中不断优化提升。【**结论**】研究成果提升了科技大数据知识发现平台以及相关情报智能分析系统的数据服务维度及深度。

关键词: 科技大数据　数据增值　丰富化方法
分类号: TP391
DOI: 10.11925/infotech.2096-3467.2018.1355

1　引　言

随着网络资源的海量膨胀，获取及组织科技大数据的方法，在信息组织、信息检索、电子商务等领域被广泛研究及应用。国内外科研人员基于智能农业[1]、智能交通[2]、智能政务[3]等专项领域进行数据增值方法及应用研究，数据抓取、实体识别、数据规范化、数据融合等数据增值方法被广泛研究与应用，但研究内容分散于各个学科。

本文就各个学科的科技文献大数据建立一套通用的增值丰富化框架体系，所覆盖的科技文献大数据主要来源于科研院所、高等院校、科技企业、政府机构所发表的科技文献，涵盖论文(期刊论文、会议论文、学位论文)、标准、专利、科技报告、图书、期刊、项目 7 类资源，数据来源包括维普科技期刊论文、中国科学院机构知识库论文、中国科学院学位论文、Web of Science 科技文献、中国国家知识产权局专利数据、中国标准化研究院的标准数据、中国国家自然科学基金会等共计 60 多种。由于科研文献中包括的科研实体字段内容有限，仅包含会议名称、项目名称、机构名称信息，丰富化建设不仅需要完成文献中科研实体的标注，也需要进行实体属性丰富化，促进文献信息的分析与展示。

对数据增值丰富化的方法研究包括多个方面。程秀峰等针对科研数据管理系统 RDM 增值服务的需求，建立融合用户需求的科研数据服务模型，从用户、学科、资源三个层面为提升 RDM 系统的感知能力提供数据支撑及优化建议[4]。于倩倩等采用第三方数据源的方式，以 NSTL 的加工规范为基础，把第三方元数据集成到 NSTL 系统中[5]。田磊通过对主题爬虫搜索策略的研究、设计与实现完成了基于 Hadoop 的大数据平台工具 MapReduce 的主题新闻获取、存档系统[6]。王颖等对以实体及实体关系为中心的语义检索发现系统进行探索[7]。孙海霞等就科技文献中机构名称规范

通讯作者：谢靖, ORCID: 0000-0001-6698-1786, E-mail: xiej@mail.las.ac.cn。
*本文系国家科技图书文献中心下一代国家科技创新开放知识服务系统项目“用户画像模型及关键技术研究”(项目编号：科 1810)和中国科学院文献情报能力建设专项项目“基于大数据计算的知识发现服务平台建设”(项目编号：院 1759)的研究成果之一。

化及匹配策略进行相应的研究[8]。刘琨等分析中国图情领域名称规范的研究性论文，指出名称规范在机构知识库、知识组织、专利情报领域方面的重要性[9]。孟小峰等对 Web 数据、科学数据、商业数据融合进行案例分析，对三种数据融合技术模式和本体对齐技术、实体链接技术、冲突解决及关系推演技术进行相应分析，提出融合技术存在跨学科、跨领域问题，跨语言、跨媒体问题[10]。

中国科学院文献情报中心作为中国科学院的科技知识资源支撑服务单位，同很多文献数据库出版商有合作关系，而且拥有丰富的自建科技资源。但已有资源的数据深度不足以支撑科研大数据分析需求，同样面临着科研大数据缺失的问题。为形成多维、深度的数据分析服务，建立基于大数据平台的科技大数据知识发现平台，采用语义检索方式提升知识发现能力及数据分析服务能力，本文通过借鉴以上科研成果丰富化建设的方法，对 5 种科研实体——科研人员、科研机构、学术期刊、学术会议、科研项目，通过采用以下方法：数据增值、数据抓取、实体识别、数据规范化、数据融合，完成对文献实体信息的补充，完成丰富化实体数据同科技大数据文献中科研实体数据的融合及实体关系的建立，为上层数据分析服务提供有力的数据支撑。

2 科技大数据增值建设框架

通过对科技文献数据的分析，确定对科研人员、科研机构、学术期刊、学术会议、科研项目 5 类科研实体进行丰富化建设。科技大数据增值建设分为三个层面进行：科研大数据数据源遴选；多来源数据的获取、清洗、规范化建设；规范化实体数据融合。科技大数据增值与丰富化建设数据的框架如图 1 所示。

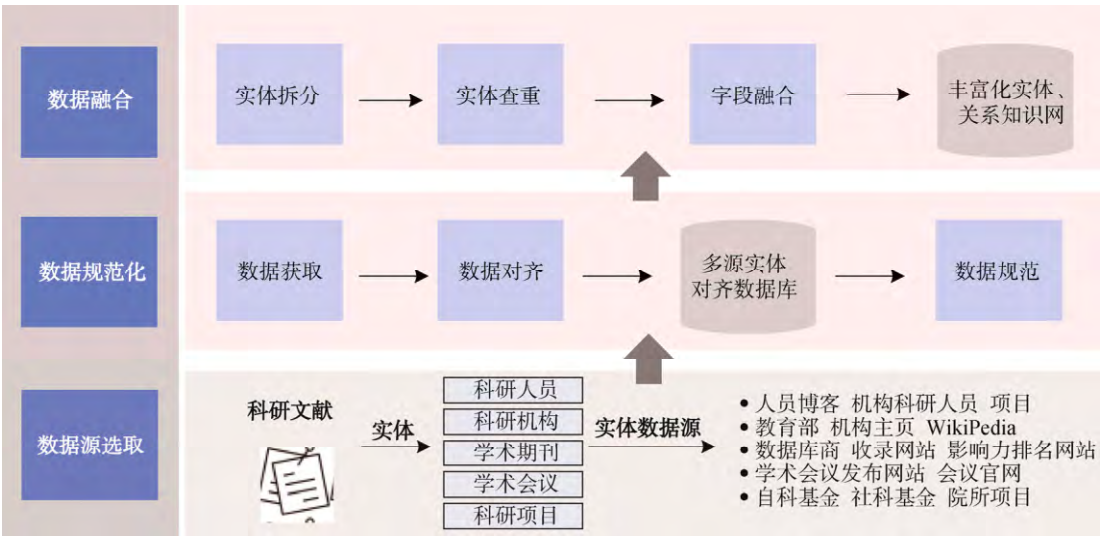

图 1 科技大数据平台丰富化建设整体框架

(1) 科研大数据数据源遴选，通过对科研实体数据增值数据源的调研与比对，为科研大数据增值建设提供一套全面、可靠的丰富化建设数据源。

(2) 科研数据的获取与规范化建设，完成科研大数据的获取以及数据内容的规范化，通过规范化建设，保证入库数据的正确性，为数据融合打下基础。

(3) 数据融合，为了解决相同实体的信息合并，构建更全面的实体数据字段及实体关系。实体数据的融合对文献数据中的实体数据形成补充，把单薄实体变成厚实体；实体关系的融合，通过对丰富化科技大数据进行关系抽取，结合科技文献中抽取的实体关系，构建出全向的科技实体关系[4,7]。

通过数据丰富化建设，可构造出拥有丰满信息的实体以及丰富的实体关系，为上层知识服务的多维数据分析服务提供丰富的分析数据[4-5,7,10]。

3 科技大数据实体数据源选取

科技实体数据源的选取非常广泛，为保证数据质

量，数据源的可靠性评测依据顺序是：数据源是否为官网；数据源的影响力；数据源的数据质量；数据获取的难易程度。分别对 5 类科研实体的数据源选取进行介绍，其中重点介绍以科研人员为例的实体数据源选取过程。

3.1　科研人员数据源选取

科研人员数据源以中国科学院的研究人员、中外高校科研人员作为目标数据，初步选取科研人员的主页、科研机构网站的科研人员介绍、科研网站上的注册科研人员信息、国内外基金项目中的科研人员信息作为丰富化数据源。

(1) 科研机构网站的科研人员信息

对中外高校机构网站科研人员进行调研，高校机构网站教师信息分布于每个学院，但由于高校下属每个学院的机构主页风格不相同，无法快速量化获取高校教师信息。

中国科学院机构网站建设比较规范统一，每所机构都对科研人员按照人员分类(研究员、副研究员、杰出青年等类型)进行个人主页建设，个人主页结构规范、信息完整，最终选取中国科学院 49 所京内研究单位、65 所京外研究单位、4 所公共支撑单位进行人员信息的获取。

(2) 机构知识库等科研人员信息

对中国科学院机构知识库 IR①、中国科讯②、中国科学家在线网站 iAuthor③、学术社交网站 ResearchGate④分别进行调研。

IR 是中国科学院兰州文献情报中心建设的网站，收集了中国科学院科研人员信息及科研成果[11]，该网站以 API 接口形式提供人员信息给中国科学院内部使用。中国科讯是中国科学院开发的资源服务 APP，包括 iOS、Android、Web 三种服务模式，其中包括大量在该平台注册的科研人员的姓名、邮箱等信息。

中国科学家在线 iAuthor 收集了 6 万多个科研人员的信息，以 API 接口的形式提供人员的姓名、机构、邮箱、研究方向等信息[12]。

ResearchGate 网站拥有 1500 万科研人员，一部分科研人员同意公开个人信息，可以直接获取该部分人员的信息，包括姓名、机构信息，但人员信息以英文为主，而且人员姓名以用户自己录入为主，存在很多不规范数据，作为一种备用数据源，供后续深入研究。

(3) 国内外基金项目中的科研人员信息

中国科学院文献情报中心的监测团队获取到 179 个国家近 580 个项目的数据，从中可以抽取项目申请人实体信息，包括科研人员姓名、机构、职称、性别、国家、研究方向等规范化信息。

另外，中国科学院文献情报中心通过同其他科研机构进行项目合作，可以获取到合作机构的人员照片及人员邮箱、研究方向等科研人员信息。

(4) 科研人员主页信息

对新浪微博等博客网站的科研人员信息进行调研，发现用户姓名存在大量别名，该类数据存在大量的噪声数据，因此未选用该类数据作为丰富化建设的数据源。通过调研，科研人员增值建设数据源选用结果如表 1 所示。

表 1　科研人员增值建设数据源选用结果

科研人员类型	调研数据源	是否采用
中国科学院科研人员	各研究所机构网站	√
	中国科学院机构知识库	√
	中国科讯注册用户	√
高校科研人员	高校官网	×
	中国科学家在线	√
其他科研人员	项目数据	√
	由研究机构提供	√
	博客网站	×

3.2　科研机构数据源选取

科研机构按照中国科学院机构、国内外高校、国内外科研机构的顺序进行数据源调研与遴选。

中国科学院各类科研机构采用中国科学院 IR 机构网站信息，其中包括 609 个中国科学院的研究所、实验室信息。国内高校列表信息用中国教育部官网①公布的中国的高等学校共计 2 914 所高校。世界知名大学数据获

①http://www.irgrid.ac.cn/.

②http://stpaper.cn.

③https://iauthor.cn.

④https://www.researchgate.net/.

取选用USNews(4大权威世界大学排名之一)公布的全球大学排名信息，排名大学分布于 74 个国家，共计 1 250 所高校。Grid 机构网站[②]包含 221 个国家近 9 万个科研机构及机构关系数据。DBpedia 中选取机构类型为大学、实验室的机构数据。高校的中英文标准名称、简称、曾用名、机构中英文介绍等信息选用百度学术、Wikipedia[③]两种数据源进行补充，机构的地理位置信息采用 Google Places 接口进行补充。

3.3 学术会议数据源选取

由于学术会议网站比较分散，对国内外学术会议网站进行调研，选用中国科学院的学术会议网站[④]、中国科学院国际会议服务平台[⑤]、中国计算机学会 CCF 分类会议、国际学术会议网站 IEEE、ACM 会议等作为会议丰富化数据源。

3.4 学术期刊数据源选取

(1) 通过对同中国科学院合作的数据库出版商如 IEEE、Springer 等进行调研，选取 80 家使用频度最高的中外文数据库商进行期刊数据的获取;

(2) 中国科学院文献情报中心的自加工期刊数据，共计有 4 万种电子期刊的加工数据;

(3) 期刊的权威评测数据[13]，包括期刊影响因子、期刊 JCR 分区、期刊学科分类(中图分类、科图分类、ESI 分类等)、期刊收录信息(北大核心、南大核心等)。

3.5 科研项目数据源选取

中国科学院文献情报中心的监测团队，自建了一套科研项目数据丰富化方法，通过对中华人民共和国科学技术部、中国国家自然科学基金委员会、美国国家自然科学基金、日本科学研究费助成事业厅、加拿大自然科学与工程研究委员会、英国科学与技术设施理事会等共计 49 种占世界项目比重较大的国家项目基金网站[14]上进行科研项目数据的获取与分析。

此外，各研究所与中国科学院文献情报中心已形成科研项目数据的合作共享，文献情报中心建立了供科研机构进行项目数据提交的平台，以文献情报中心要求的标准格式提交的项目数据可以直接进入到中国科学院文献情报中心的科技大数据知识发现平台上供用户使用。

4 实体数据获取及对齐方法

从增值数据源获取的实体数据，需进行实体数据的对齐、入库、规范化，为实体数据融合提供规范化数据内容。本文参照NSTL统一文献元数据标准[15]，结合各类实体数据获取网站的内容字段，构建出一套针对5种实体数据的对齐格式，选用MySQL作为增值数据存档数据库，并且制定出一套实体字段规范规则。

实体对齐用于判断不同信息来源的实体是否指向真实世界中的同一对象[16]，本研究采用的实体及属性对齐方法共经过三个步骤。

(1) 不同来源的实体属性字段映射到统一的实体数据库表字段;

(2) 对实体属性内容进行清洗规范化;

(3) 选用实体字段信息组合作为实体唯一标识生成依据，选定实体查重字段及字段查重顺序，完成实体数据的对齐。

4.1 实体数据获取方法

根据科研实体数据的数据源提供方式，采用以下两种方式完成丰富化实体数据的获取:

(1) 以API形式提供的数据，如: 科研人员数据源 IR、iAuthor、机构位置数据 Google Places 等，通过开发接口数据调用代码，完成接口数据的获取;

(2) 网站内容采集数据[6]，采用 WebCollector、JSoup、BeautifulSoup 等工具完成静态、动态网页信息的解析及结构化数据的获取。

通过数据抓取、数据字段对齐，构建完成多来源实体对齐数据库。

4.2 实体属性字段对齐方法

科研大数据每种实体选用多种数据源进行数据增值，NSTL 统一文献元数据标准[15]针对贡献者/机构、会议、基金等 13 种元素集的字段内容进行相应规范，

①http://www.moe.gov.cn/.

②https://www.grid.ac/.

③https://en.wikipedia.org.

④http://www.cas.cn/xs/index.shtml.

⑤http://csp.escience.cn.

并对实体对象之间的关系进行概括：组成关系、相关关系、规范关系、沿革关系、引用关系。本文依照 NSTL 统一文献元数据标准提炼出的实体字段为基础，针对科研人员[17]、科研机构等 5 种科研实体，设计出每种科研实体的数据字段表。

以科研人员为例，通过对科研人员数据源字段调研，最终确定科研人员多数据源字段对齐格式如表 2 所示，其中用于实体唯一标识的字段包括 ORCID、邮箱、姓名、机构。数据属性对齐建设共计完成科研人员 21 种对齐字段、科研机构 20 种对齐字段、期刊 36 种对齐字段、学术会议 20 种对齐字段的设计。

4.3　实体属性规范化方法

通过对实体规范化方法的调研[9]，并且为实现实体的快速规范化，针对实体属性字段内容采用不同的规范方法，形成实体数据规范化流程，如图 2 所示，规范化过程包括：非法内容过滤，对不可显示字符等非法内容进行过滤；字段标准化，对实体中机构名称采用中国科学院机构名称规范库进行标准化、机构英文名称汉化、人名拼音变体等；实体字段合法性检查，如实体中的链接信息，采用链接规则对链接合法性进行检测，规范后通过合法性检测的字段才能作为实体属性字段融合的内容。

表 2　科研人员多数据源字段对齐格式

对齐字段	对齐字段	对齐字段
姓名	职称	邮编
机构	学术头衔	简历
邮箱	专业	研究领域
部门、院系	学历	荣誉
实验室、研究组	电话	个人主页
性别	传真	用户照片
行政职务	通讯地址	ORCID

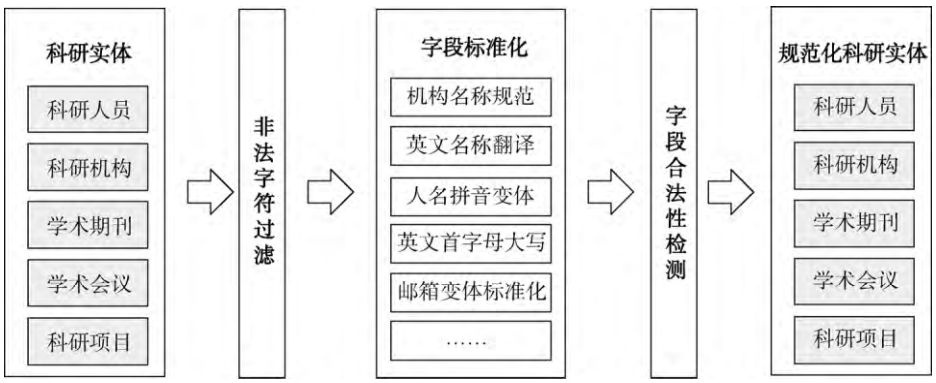

图 2　实体数据规范化方法

4.4　丰富化实体及属性融合方法

多来源实体数据存在很多重复实体，虽然前期数据采集完成了数据字段的对齐，但每种数据源的实体字段数、字段内容存在差异，需要对实体信息、实体关系进行融合建设，对不同来源的同一实体完成实体属性值的融合。实体数据对齐顺序如下。

(1) 优先选用官网数据作为数据源。官网数据字段相对完整，数据内容也是最新的，如中国科学院的科研人员，优先选用科研人员所在院所主页上的信息作为数据源，高等院校的科研人员优先选用学校学院下该科研人员的信息作为补充数据源；

(2) 选用权威数据知识库网站等作为数据补充，数据字段相对完善，数据内容的更新相对滞后，如中国科学院机构知识库网格 IR GRID[11]的数据及中国科学家在线 iAuthor[12]的数据；

(3) 选用质量较好的科研文献中的结构化数据作为数据补充源，数据字段内容相对较少，数据内容存在时间跨度，如各个国家的基金项目、专利数据中的科研人员数据信息。

实体数据融合建设流程如图 3 所示，采用 ElasticSearch(ES)分布式全文搜索引擎作为数据融合平台，从多来源实体对齐数据库中读取每种实体数据信息，完成实体增值数据的分离，如增值科研人员信息中包含机构信息，可分离成人员、机构两种补充数据，补充数据依照大数据知识发现平台的实体及关系入库规则完成实体及实体关系的融合。

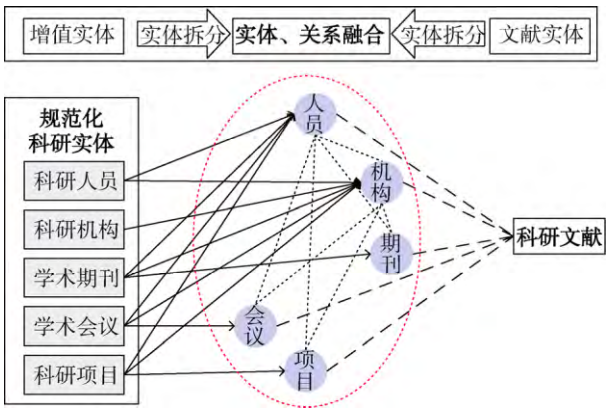

图 3　实体及实体关系融合

为完成不同来源的丰富化实体及科研文献中的实体对齐，采用两种方式完成实体对齐。

(1) 构建每种实体唯一标识符 UUID 的生成规则，包括生成 UUID 所使用的字段及字段顺序;

(2) 限定实体查重字段、字段变体、字段顺序。

以科研人员实体为样例，科研人员 UUID 的生成字段及顺序是: ORCID、邮箱、用户名、用户机构。新录入实体同 ES 中已有实体的查重依次选用属性字段顺序为: ORCID 邮箱、用户名、用户机构名称，中文用户名需协同使用姓名的各类拼音变体，机构名采用机构的中英文标准名称。通过上述三类实体属性字段组合依次检索，如果查询到相同的科研人员，完成人员信息的合并，否则该科研人员被认定为新的科研人员，完成实体信息的录入。

4.5　实体规范化及实体对齐

实体规范化的目的是保证实体的对齐，完成不同来源的实体融合。丰富化实体在规范化过程中遇到很多问题，关键问题有以下 4 种。

(1) 实体字段内容不规范，需建立特殊字符过滤函数，对不可见字符、空字符等进行过滤;

(2) 实体字段名称存在多种变体，如中文人名“罗平”，在其发表的英文名称里有“Luo Ping”、“Ping Luo”等，采用人名拼音转换函数完成该人员规范化姓名的补充;

(3) 实体字段多语种，由于丰富化实体数据来源包含国内外，而当前服务用户重点面向中国的用户，当前采用百度翻译、谷歌翻译 API 接口完成国外实体数据查重字段的翻译，转换实体属性为中文，保障中英文实体对齐与信息互补;

(4) 实体字段表示方式不规范，如人员的职称是正高，标准表示方式为正高级工程师，该字段内容的表示方式有限，通过对丰富化数据库字段的内容进行聚类分析，构建非规范化字段到规范化字段的映射规则文件，采用规则匹配完成目标实体属性的规范。

科研实体对齐是为了完成科研同一科研实体的字段合并。科研实体对齐的关键问题如下。

(1) 实体名称中英文变体对应，从标准后的丰富化数据中，建立实体中英文名称对应映射关系，如果当前实体名称无法同已有实体匹配，转换当前实体名称为其他语种的规范化表示方式继续进行实体匹配查询，如科研人员罗平，中国科学院文献情报中心，其英文实体表示方式可能是 Luo Ping, National Science Library, Chinese Academy of Sciences，通过实体属性多种语种类型组合循环匹配查找，达到实体对齐;

(2) 实体歧义，同一科研实体由于标点、大小写等无法对齐，建立公用的正则过滤函数，过滤后的实体属性只保留汉字及字母，采用过滤后的内容进行匹配，如同一篇论文，由于收录来源不同，文献标题中的标点等导致文献无法匹配，采用该过滤规则，可完成实体的匹配;

(3) 实体属性内容选用标准，实体属性在 ES 中采用列表的存档方式，保障实体属性内容可被覆盖及追加融合，高质量来源的数据源优先入库，同时以重写覆盖方式导入到分面字段中，后续来源的实体属性存档于该属性非分面字段列表中;

(4) 数据规模大，实体匹配效率问题，选用 ES 分布式索引方式，保障实体数据的查重及录入速度，另外基于 4.4 节的三种实体匹配规则，保障大数据量实体的快速匹配。

5　科研人员增值数据建设示例

5.1　科研人员增值数据获取

科研人员增值数据建设方法有通过 API 方式获取及通过网络采集方式获取补充数据，本文介绍网络采集方式获取科研人员增值数据，数据获取流程如图 4 所示。

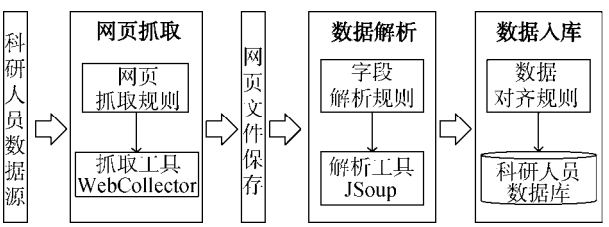

图 4　科研人员数据获取

根据数据源，采用 WebCollector 网页数据抓取工具进行抓取，需要对每个机构的人员配置抓取规则，规则包括人员网页抓取入口、抓取深度、排除链接等规则信息，抓取到的网页存档到对应机构目录下，采用 JSoup 网页解析工具，结合人员网页数据字段解析规则，完成科研人员字段解析，人员信息解析结果采用对齐规则入到科研人员数据库。

5.2　科研人员增值数据规范化

以科研人员中部分实体字段规范化为例，选用部分字段规范化规则进行介绍，如图 5 所示。

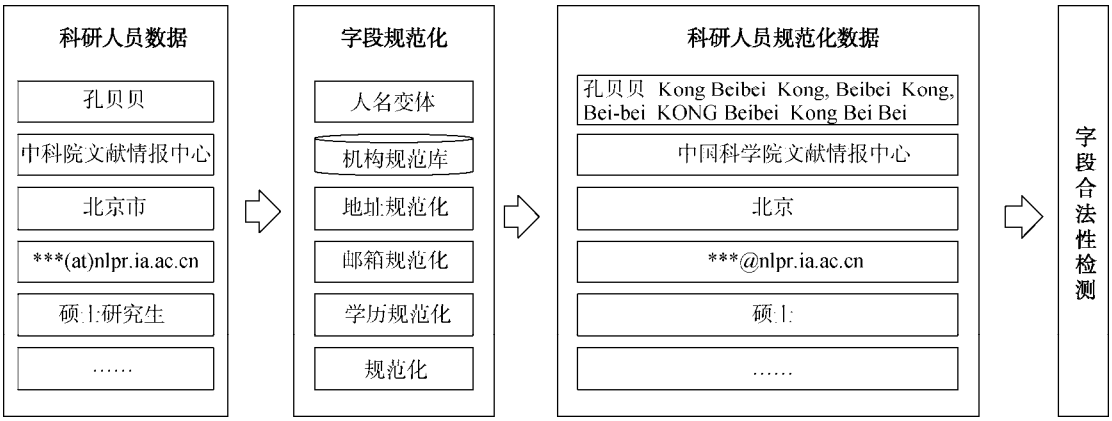

图 5　科研人员规范化

(1) 人员中文名称，采用名称变体规则，变为多种拼音表示方式，促使中文作者同英文作者融合;

(2) 机构名称“中科院文献情报中心”经机构规范化库标准化为中英文两种名称: “中国科学院文献情报中心”、“National Science Library, Chinese Academy of Sciences”;

(3) 省份及城市信息“北京市”规范化为“北京”;

(4) 邮箱信息“***(at)nlpr.ia.ac.cn”规范化为“***@nlpr.ia.ac.cn”。

每类实体中的同一类字段采用相同的规范化规则转换为可供融合使用的标准字段内容。规范化后的数据，采用校验规则对字段内容合法性进行检测，校验合法的字段供融合查重字段及融合字段内容使用。

5.3　科研人员增值数据融合

图 5 中的科研人员数据，通过实体分离，构建出科研人员孔贝贝的属性信息、科研机构中国科学院文献情报中心两种不同的实体。样例中科研人员实体依次选用两类信息按照 4.4 节的查重顺序进行查重:

(1) 邮箱: ***@nlpr.ia.ac.cn;

(2) 姓名协同机构名称: 孔贝贝、中国科学院文献情报中心的各类变体。

科研机构实体分离出的信息包括机构名称、机构城市、机构地址、机构研究方向信息，查重依次使用字段为:

(1) 机构中英文名称及机构城市，如: 中国科学院文献情报中心、北京，如果查询不到再采用机构的英文信息 National Science Library, Chinese Academy of Sciences 及 Beijing;

(2) 如果以上查询不到，采用机构中文英名称及所属国家的中英文组合进行查重，如: 中国科学院文献情报中心、中国。

明确了实体对齐选用字段属性，再依据 4.4 节丰富化实体及属性融合方法，明确科研人员的数据源入库顺序，优质数据优先入库，质量差的数据作为补充数据后入。科研人员数据源入库顺序依次为: 中国科学院各研究所机构网站; 中国科学家在线; 中国科讯注册用户; 中国科学院机构知识库; 项目分离出的科研人员; 研究机构提供科研人员。

通过实体拆分及实体字段融合完成该科研人员、科研人员所在机构数据补充，同时完成研究人员同该科研机构关系的建立。

6　科技大数据丰富化工具建设及建设效果

6.1　科研大数据丰富化工具建设

科技大数据丰富化建设中完成了 4 类工具。

(1) 增值数据抓取工具，针对科研人员、科研机

构、学术会议、学术期刊实体采用 Java 及 Python 两种开发语言，基于 WebCollector、JSoup、BeautifulSoup 等数据抓取及解析工具;

(2) 数据补充接口工具，期刊的收录、影响因子、学科分类信息作为公用数据供大数据服务使用，基于 Spring Boot 框架，采用 Redis 缓存机制，完成 28 种数据补充工具的开发;

(3) 数据清洗及规范化工具，构造了数据过滤及每种数据类型的公用规范化工具;

(4) 实体及实体关系融合工具，基于 ES 分布式索引，通过增值实体数据的拆分、实体数据及实体关系融合规则构建数据融合工具。

6.2 科研大数据丰富化建设效果

通过科技大数据增值建设，共计完成 186 万个国内外科研人员，4.5 万种中外文期刊；中国大学、国内外的教育机构、公司、医院等共计 19 万个机构；涉及计算机、医学、教育等学科的近 8 万条会议数据的增值，经过增值建设，实体字段信息扩展为原来的 5-10 倍左右。

科技大数据增值建设数据在科技大数据知识发现平台(慧眼)①、智能随身科研助理(慧科研)②、NSTL 的用户画像项目等项目实际使用。增值数据，为以上平台提供科研实体详情信息，提升了科研数据的分析维度，如科研人员发表文献的影响因子、JCR 分区展示及演变等；为科研人员的画像计算提供数据基础；科研人员画像计算结果为慧科研平台的科研人员资源精准获取及精准推荐提供了数据基础。

7 结 语

科技大数据未进行增值丰富化之前，无法满足知识检索与情报分析的特殊数据分析需求，而经过科技大数据实体及实体关系的增值建设，使科研数据更加丰富，让用户通过科技大数据服务平台获取详实、精确的知识服务，也为计算型科技知识服务的发展提供了大数据基础。

本研究采用基于规则的方式实现实体数据规范化与实体信息快速融合，为保障实体数据的及时性及准确性，需进一步的研究数据更新机制；为提升科研实体的对齐效果，需对机器学习、神经网络[6,10,16,18]计算方法、Word2Vec 等实体相似度计算方法、实体链接预测技术进行研究与应用。面向多类型的科技情报服务需求，未来会持续对丰富化科研实体数据进行充实与完善，并对先进的实体规范化算法及对齐方法进行研究与大规模使用，进一步提升实体数据规范化的精准度、提升实体数据的融合效果。

参考文献:

[1] 倪芳，曾辉，卓辉，等. Web 服务在多源异构农业数据融合上的应用研究[J]. 计算机技术与发展，2016，26(8): 129-133. (Ni Fang, Zeng Hui, Zhuo Hui, et al. Research on Application of Web Services in Multi-Source Heterogeneous Data Integration on Agriculture[J]. Computer Technology and Development, 2016, 26(8): 129-133.)

[2] 陆百川，舒芹，马广露. 基于多源交通数据融合的短时交通流预测[J]. 重庆交通大学学报：自然科学版，2019，38(5): 13-19, 56. (Lu Baichuan, Shu Qin, Ma Guanglu. Short-term Traffic Flow Forecasting Based on Multi-source Traffic Data Fusion[J]. Journal of Chongqing Jiaotong University: Natural Science, 2019, 38(5): 13-19, 56.)

[3] 张卫东，左娜，陆璐. 政府网站信息资源知识融合体系架构设计[J]. 图书情报工作，2018，62(17): 112-119. (Zhang Weidong, Zuo Na, Lu Lu. Knowledge Fusion System Architecture Design of Government Website Information Resources[J]. Library and Information Service, 2018, 62(17): 112-119.)

[4] 程秀峰，王雪杰，夏立新. 科研数据管理系统中增值服务调查研究[J]. 情报科学，2018，36(10): 77-83. (Cheng Xiufeng, Wang Xuejie, Xia Lixin. Investigation on Value-added Service in Research Data Management Systems[J]. Information Science, 2018, 36(10): 77-83.)

[5] 于倩倩，张建勇. NSTL 集成利用第三方来源元数据的实践与探索[J]. 现代图书情报技术，2016(1): 97-102. (Yu Qianqian, Zhang Jianyong. Practices of NSTL Integrating and Using Third-party Metadata[J]. New Technology of Library and Information Service, 2016(1): 97-102.)

[6] 田磊. 主题爬虫搜索策略的设计与实现[D]. 北京：北京邮电大学，2017. (Tian Lei. Research and Implementation of

①http://scholareye.cn.

②http://scholarin.cn.

Focused Crawler with Search Strategy[D]. Beijing: Beijing University of Posts and Telecommunications, 2017.)

[7] 王颖, 吴振新, 谢靖. 面向科技文献的语义检索系统研究综述[J]. 现代图书情报技术, 2015(5): 1-7. (Wang Ying, Wu Zhenxin, Xie Jing. Review on Semantic Retrieval System for Scientific Literature[J]. New Technology of Library and Information Service, 2015(5): 1-7.)

[8] 孙海霞, 王蕾, 吴英杰, 等. 科技文献数据库中机构名称匹配策略研究[J]. 数据分析与知识发现, 2018, 2(8): 88-97. (Sun Haixia, Wang Lei, Wu Yingjie, et al. Matching Strategies for Institution Names in Literature Database[J]. Data Analysis and Knowledge Discovery, 2018, 2(8): 88-97.)

[9] 刘琨, 李春利, 白福春. 我国图情领域名称规范文献计量研究[J]. 图书馆工作与研究, 2017(12): 66-71. (Liu Kun, Li Chunli, Bai Fuchun. Biobiometric Study on the Name Authority Literatures in Library and Information Field in China[J]. Library Work and Study, 2017(12): 66-71.)

[10] 孟小峰, 杜治娟. 大数据融合研究: 问题与挑战[J].计算机研究与发展, 2016, 53(2): 231-246. (Meng Xiaofeng, Du Zhijuan. Research on the Big Data Fusion: Issues and Challenges[J]. Journal of Computer Research and Development, 2016, 53(2): 231-246.)

[11] Zhu Z, Zhang D, Li L, et al. Developing Institutional Repositories Network: Taking IR Grid at Chinese Academy of Sciences as an Example[J]. Chinese Journal of Library and Information Science, 2011, 4(Z1): 24-34.

[12] 张建勇, 黄永文, 于倩倩, 等. 中国 ORCID 注册平台 iAuthor 的设计与实现[J]. 现代图书情报技术, 2015(3): 84-91. (Zhang Jianyong, Huang Yongwen, Yu Qianqian, et al. Design and Implementation of ORCID China Service 'iAuthor'[J]. New Technology of Library and Information Service, 2015(3): 84-91.)

[13] Vidal-Infer A, Tarazona B, Alonso-Arroyo A, et al. Public Availability of Research Data in Dentistry Journals Indexed in Journal Citation Reports[J]. Clinical Oral Investigations, 2018, 22(1): 275-280.

[14] 张璐杰. 国家自然科学基金项目立项同行评议质量控制研究[D]. 北京: 北京科技大学, 2015. (Zhang Lujie. Research on the Quality Controlment of Peer Review About NSFC Project Set-up[D]. Beijing: University of Science and Technology Beijing, 2015.)

[15] 张建勇, 于倩倩, 黄永文, 等. NSTL 统一文献元数据标准的设计与思考[J]. 数字图书馆论坛, 2016(2): 33-38. (Zhang Jianyong, Yu Qianqian, Huang Yongwen, et al. Metadata Standard Design of NSTL Unified Literature[J]. Digital Library Forum, 2016(2): 33-38.)

[16] 杨秀璋. 实体和属性对齐方法的研究与实现[D]. 北京: 北京理工大学, 2016. (Yang Xiuzhang. Research and Implementation on Entity Alignment and Attribute Alignment[D]. Beijing: Beijing Institute of Technology, 2016.)

[17] 任平. 高校教师个人信息数据融合的研究[D]. 北京: 北京交通大学, 2017. (Ren Ping. Research on Data Fusion of Personal Information in Colleges and Universities[D]. Beijing: Beijing Jiaotong University, 2017.)

[18] 张琳, 秦策, 叶文豪. 基于条件随机场的法言法语实体自动识别模型研究[J]. 数据分析与知识发现, 2017, 1(11): 46-52. (Zhang Lin, Qin Ce, Ye Wenhao. Automatic Recognition of Legal Language Entities Based on Conditional Random Fields[J]. Data Analysis and Knowledge Discovery, 2017, 1(11): 46-52.)

作者贡献声明:

孔贝贝: 科研数据增值方法研究, 数据增值工具开发, 数据采集与清洗, 论文主要撰写人;
谢靖: 科研数据丰富化方案及架构设计, 协同撰写论文;
钱力: 科研数据丰富化方案设计及实体规范规则设计, 论文修订;
常志军: 科研数据采集, 实体数据对齐规则协同制定;
吴振新: 科研数据采集, 相关文献调研。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: kongbb@mail.las.ac.cn。
[1] 孔贝贝, 谢靖, 钱力, 常志军, 吴振新. 丰富化实体数据对齐字段.doc. 科技大数据增值丰富化方法研究与工具研发实体数据对齐字段.
[2] 孔贝贝, 谢靖, 钱力, 常志军, 吴振新. 基于丰富化数据的数据补充接口. 28 种补充数据接口工具.
[3] 孔贝贝, 谢靖, 钱力, 常志军, 吴振新. Util.java. 数据公共规范化工具.
[4] 孔贝贝, 谢靖, 钱力, 常志军, 吴振新. CommonEsInfo.java. 实体及实体关系融合方法工具.

收稿日期: 2018-12-03
收修改稿日期: 2019-03-26

Methodology and Tools to Enrich Sci-Tech Big Data

Kong Beibei[1] Xie Jing[1,2] Qian Li[1,2] Chang Zhijun[1,2] Wu Zhenxin[1,2]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: [Objective] This paper tries to address the issues facing sci-tech big data, such as source dispersal, low quality, and poor content. **[Methods]** We used value-added computing methods, such as data cleansing, entity alignment, entity field fusion, conflict detection, etc., to develop tools for the enrichment of sci-tech big data. **[Results]** The developed tools achieved entity data alignment at the levels of personnel, organization, conference, journal and relationship among them. The contents of the entity fields were increased by 5 to 10 times, and the entity analysis dimension was increased by 2 to 3 times. **[Limitations]** The timeliness and standardization of value-added data need to be optimized and improved based on service needs. **[Conclusions]** The proposed methods and tools enhance the knowledge discovery of the sci-tech big data and intelligent information analysis systems.

Keywords: Sci-Tech Big Data Data Appreciation Enrichment Method

精通技术的人更容易信任数字医生

宾州州立大学最近的一项研究表明，对机器性能和自身的技术能力充满信心的人更有可能接受和使用数字医疗保健服务。

“医疗领域正在广泛使用自动化系统。”研究人员指出，“我们调查了用户对这些‘机器人接待员’、‘自动护士’和‘自动医生’的接受程度，并测试了这些角色所采用的形式——人类形式、虚拟现实形式或机器人形式——是否影响用户的接受程度。”

研究人员认为，人们越来越依赖自动化系统，医疗保健行业可以从中受益。“医生的精力是有限的，他们的经验和知识也是有限的。”研究人员认为：“相比之下，机器则可以被编程为‘思考’患者症状可能指向的所有可能条件，并且机器永远不会感到疲倦。要做到这样，显然需要较高程度的自动化。”

研究人员衡量了参与者对机器的先入为主的信念和态度，即所谓的“机器启发式”，其反映了人们对机器的刻板印象，包括他们对机器的无差错、客观性和效率的信念。研究人员向参与者展示了医疗服务提供者的各种组合(如接待员、护士和医生)以及代理人类型的不同组合(如人、虚拟现实和机器)。通过与各种类型的头像进行在线聊天互动，测试参与者对这些医疗服务提供者的接受程度以及他们将来使用这些服务提供者的意图。

“我们发现，机器启发式的信念越强，对代理人的态度越积极，在将来使用服务的意图就越大。”研究人员表示，“计算机水平和对数字医疗保健提供商的接受度正相关。也就是说，计算机水平高的用户比计算机水平差的用户更容易接受机器人医生。”

此外，研究人员还注意到这两个因素的“双剂量”效应。“如果你在机器启发式的信念强而且有较高的计算机水平，那么你对自动化医疗服务提供商的态度最为积极。”在所有实验条件下，效果都是类似的。换句话说，那些高度依赖机器启发式并且计算机水平高的用户对所有形式的数字医疗保健提供者——无论是人类、虚拟现实还是机器人——几乎同样持积极态度。

研究结果表明，在医疗保健机构实施自动化的关键之一是交互界面的设计，需要能吸引对机器能力有高度信任且计算机水平较高的用户，关键之二在于聊天等功能的改进，无需将资源耗费在将医疗保健机器人的外形拟人化上。此外，提高用户的计算机水平，加强用户对机器的信赖度，能增加自动化服务的采用率。

(编译自：https://www.sciencedaily.com/releases/2019/05/190510113807.htm)

(本刊讯)